CLUSTERING HYPERTEXT WITH APPLICATIONS TO WEB SEARCHING

ABSTRACT

A method and structure for providing a database of documents comprising performing a search of the database using a query to produce query result documents, constructing a word dictionary of words within the query result documents, pruning function words from the word dictionary, forming first vectors for words remaining in a word dictionary, constructing an out-link dictionary of documents within the database that are pointed to by the query result documents, adding the query result documents to the out-link dictionary, pruning documents from the out-link dictionary that are pointed to by fewer than a first predetermined number of the query result documents, forming second vectors for documents remaining in the out-link dictionary, constructing an in-link dictionary of documents within the database that point to the query result documents, adding the query result documents to the in-link dictionary, pruning documents from the in-link dictionary that point to fewer than a second predetermined number of the query result documents, forming third vectors for documents remaining in the in-link dictionary, normalizing the first vectors, the second vectors, and the third vectors to create vector triplets for document

remaining in the in-link dictionary and the out-link dictionary, clustering the vector triplets using the *toric k-means* process, and annotating/summarizing the obtained clusters using nuggets of information, the nuggets including summary, breakthrough, review, keyword, citation, and reference.